Synopsis for Registration as Research Student

For

Doctoral Degree

On

Breast Cancer Risk Prediction and Analysis Linked to Health History, Records and Mammograms

By

Mr. Aymen Muslih Musead Al Hejri

Research Scholar

Supervisor

Dr. A. H. Sable

Assistant Professor

School of Computational Sciences, S.R.T.M University,

Nanded-431606. (MS.) India.

- Introduction

Cancer is a disease that occurs when abnormal cells grow uncontrollably, disregarding the normal rules of cell division, resulting in uncontrolled growth and proliferation of the abnormal cells. If the proliferation is allowed to continue and spread in such a way that metastasis forms, this can be fatal. If the tumor invades surrounding tissues or spreads to other parts of the body, it is referred to as malignant or cancerous [1]. Breast cancer is one of the leading causes of death among women between the ages of 20 and 59 worldwide and begins in the ducts that bring milk to the nipple or in the glands that produce breast milk. Uncontrolled growth of breast cells causes a lump that can be felt or identified by mammography [2]. According to the American Cancer Society's forecasts for 2021, there will be about 281,550 new cases in the United States of invasive breast cancer diagnosed in women, and about 43,600 death cases from breast cancer [3].

So, researchers are trying to early predict breast cancer by image analysis techniques to study mammographic images in breast cancer risk. Some researchers conducted to predict of breast cancer depends on analysis data of factors of breast cancer. In this proposal we will collect recent datasets from standard websites and collect new mammograms datasets with some causes and factors of its occurrence, then apply Image processing, Computer Vision, Machine learning, and Deep Learning methods to build a robust model capable of detecting breast cancer for both datasets, then compare results of accuracy between mammograms only and mammograms with the causes and factors affecting the occurrence of breast cancer.

2- Related Work

Breast cancer is a complicated disease that is affected by both environmental and genetic factors. Modifiable environmental variables include menarche age, parity, age at first birth, breastfeeding, BMI, menopausal hormone therapy (MHT), and alcohol consumption. Furthermore, gene mutations with a

high/moderate chance and factors such as Glucose, Insulin, HOMA, Leptin, Adiponectin, Resisting, and MCP1.

Kapoor, Pooja Middha, et al. [4] have assessed the potential impact of 13 identified breast cancer risk factors using 28 176 cases, and 32 209 controls genotyped and applied unconditional logistic regression and likelihood ratio tests for overall breast cancer risk and Bayesian false discovery probability. The number of cases and controls with data for each risk factor ranged from 23 755 cases and 30 153 controls for parity to 5078 cases and 6867 controls.

Hamed, Ghada, et al. [5] have applied a comparative study to various publicly accessible datasets, such as DDSM, CBIS-DDSM, and breast. Also, conduct comparative research of different machine learning-based detection and classification methods based on AlexNet, GoogleNet, YOLO, and RetinaNet, with YOLO and RetinaNet achieving the highest accuracy. The detection accuracy is 99.7% and 98.96%, respectively.

Devi, V. Ajantha, and Anand Nayyar.[6] have conducted a study on various types of cancer and analyzing with diagnosis based on medical image (mammograms) then applied Deep learning (CNN), the dataset consists of 569 examples, 357 benign and 212 malignant, the built model is able to identify the extracted features from mammograms as benign or malignant. The overall performance was as the following accuracy result 91.6% and 91.9, %respectively.

Basarslan, M. Sinan, and F. Kayaalp.[7] applied Machine learning and deep learning algorithms to the Coimbra dataset from the UCI repository, which had ten quantitative attributes. For classification, the applied algorithms are Support Vector Machines (SVM), Random Forest (RF), Recurrent Neural Network (RNN), and Convolutional Neural Network (CNN) were used. According to these findings, RNN outperformed the others with a score of 92 %Accuracy on both datasets.

Patrício, Miguel, et al. [8] Apply machine learning algorithms (logistic regression, random forests, support vector machines) on various clinical features, including age, BMI, Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin, and MCP-1. The presence of breast cancer in women was detected using SVM, with sensitivity ranging from 82 to 88 %and specificity ranging from 85 %to 90 p%. The Area under the curve had a 95% confidence interval of [0.87, 0.91].

Shen, Li et al. [9] proposed a model of Deep Learning assisted Efficient Adaboost Algorithm (DLA-EABA) for breast cancer detection on a variety of imaging modalities, including MRI, ultrasound, optical breast tomosynthesis, and mammography, with high accuracy levels of 97.2 %, the sensitivity of 98.3 per cent, and specificity of 96.5 %.

Sha, Zijun, Lin Hu, and Babak Daneshvar Rouyendegh. [10] proposed a method to find the cancerous Area in a mammogram image using noise reduction, optimal image segmentation based on the convolutional neural network, and optimized feature extraction and feature selection based on the grasshopper optimization algorithm. The total number of images in the dataset was 2942 (322 MIAS database and 2620 DDSM), compared using ten different methods. Ninety-six per cent sensitivity, 93 %precision, 85 %PPV, 97 %NPV, and 92 %accuracy.

Saber, Abeer, et al.[11] developed a deep learning (DL) model using transfer learning (TL). Inception V3, ResNet50, Visual Geometry Group Networks (VGG)-19, VGG-16, and Inception-V2 ResNet are examples of pre-trained convolutional neural networks (CNNs). The suggested model's performance was evaluated using six metrics: accuracy, sensitivity, specificity, precision, F-score, and AUC. According to the results, the VGG16 model's TL is effective for BC diagnosis by classifying mammography breast images with 98.96% accuracy for 80-20 method and 98.87% accuracy for 10-fold cross-validation method.

3- Problem Statement

Breast cancer is one of the most common causes of death among women, so doctors and researchers in machine learning and deep learning seek early detection of breast cancer cases. Current studies focus on the classification and detection aspect of a pre-existing dataset, such as DDSM Dataset, CBIS-DDSM, MIAS Dataset, BCDR Dataset and INbreast Dataset with different algorithms and techniques, with varying results of accuracy, but without taking into account the analysis of carcinogens as part of the learning data. However, some studies are limited to only some factors without analysis of Mammogram images

Therefore, in this proposal, a model will be built by combining mammography with an analysis of the etiology of cancer factors for each patient and producing a complete classification and detection model using modern deep learning algorithms.

4- Objectives

-To collect mammograms from many resources' standard web sites.

-To Build mammograms from local hospitals with causes of breast cancer for each patient who has mammograms.

-To Extract and analyses features from both datasets.

-To combine features of both datasets and build machine learning and deep learning models to detect and predict breast cancer.

- To compare the different models and select the high Accuracy Model for classification and diagnosis.

5- Research Project Methodology, Methods, Design & Materials Needed

As shown in figure (1), this proposed method describes the process of the machine and deep learning models for breast cancer detection and prediction. The process started from data collection, then pre-processing, and Feature extraction, then applied the training and testing the model. The result will show whether the patient is infected or not.

First Phase: The Data of mammograms will be collected directly from TATA hospital in Mumbai with questionnaires about causes to build.

Second Phase: Acquisition of mammograms and set film format. Also distribute google form of the causes to be filled for each case.

Third Phase: Data pre-processing.

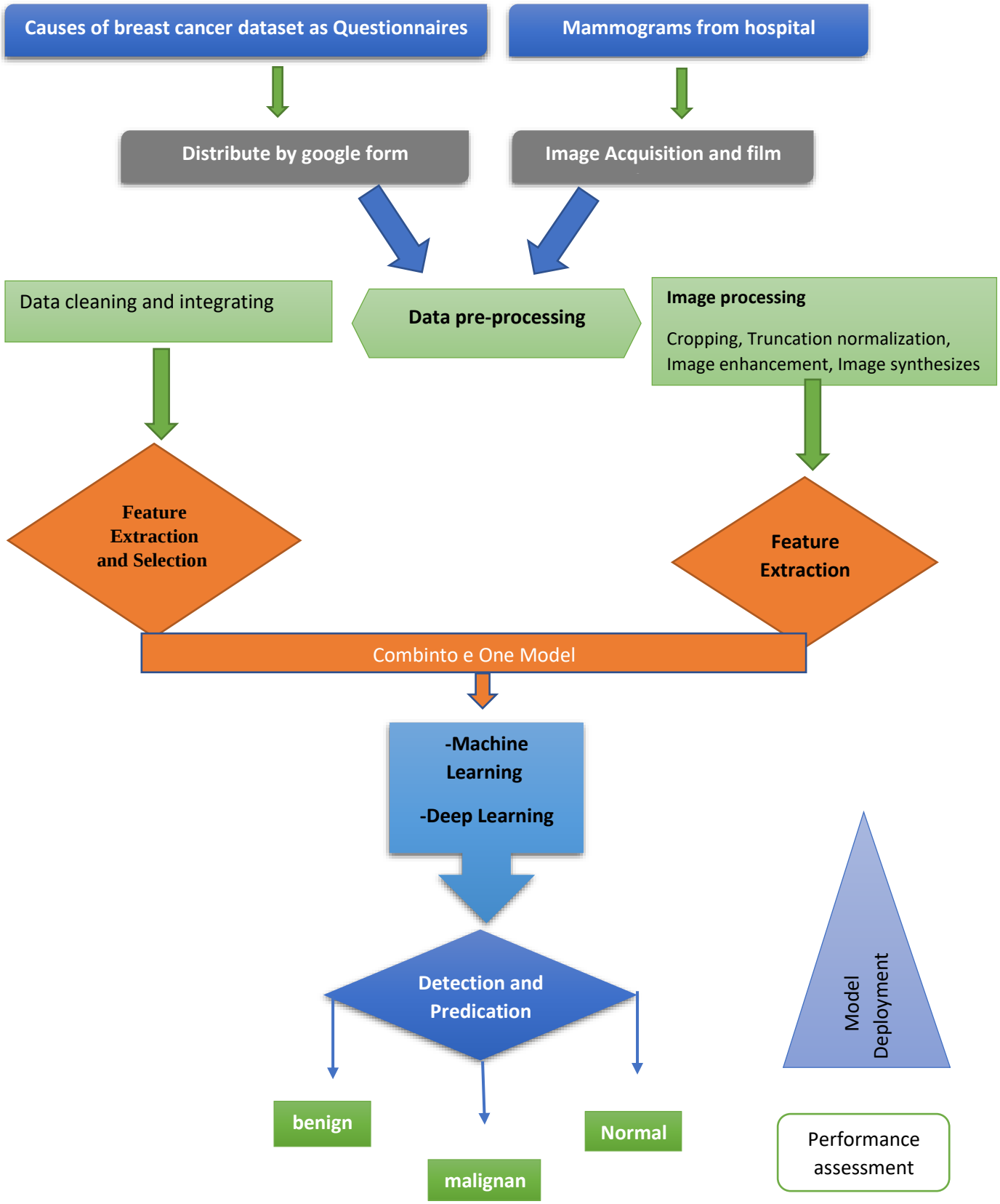
Fourth Phase: Feature extraction.

Fifth Phase: apply Machine learning and deep learning to find High accuracy model, deploy it and build interface.

Materials that will be used

Software: MATLAB and python, Spyder IDE, clinical aspects of breast cancer detection and diagnostics

Hardware: Laptop, NVIDIA card graphic.


Causes of breast cancer dataset as Questionnaires
Mammograms from hospital
Distribute by google form
Image Acquisition and film
Data cleaning and integrating
Data pre-processing
Image processing
Cropping, Truncation normalization, Image enhancement, Image synthesizes
Feature Extraction and Selection
Feature Extraction
Combinto e One Model
-Machine Learning
-Deep Learning
Detection and Predication
benign
malignan
Normal
Model Deployment
Performance assessment

6- Scope of work

The approximate duration proposed to complete research project in four years. Presentation will be done after every 6 months

First year: Literature review and collect new dataset with causes.

Second year: Implementation of basic preprocessing algorithms to improve the image quality, segmentation and deter size of cancer.

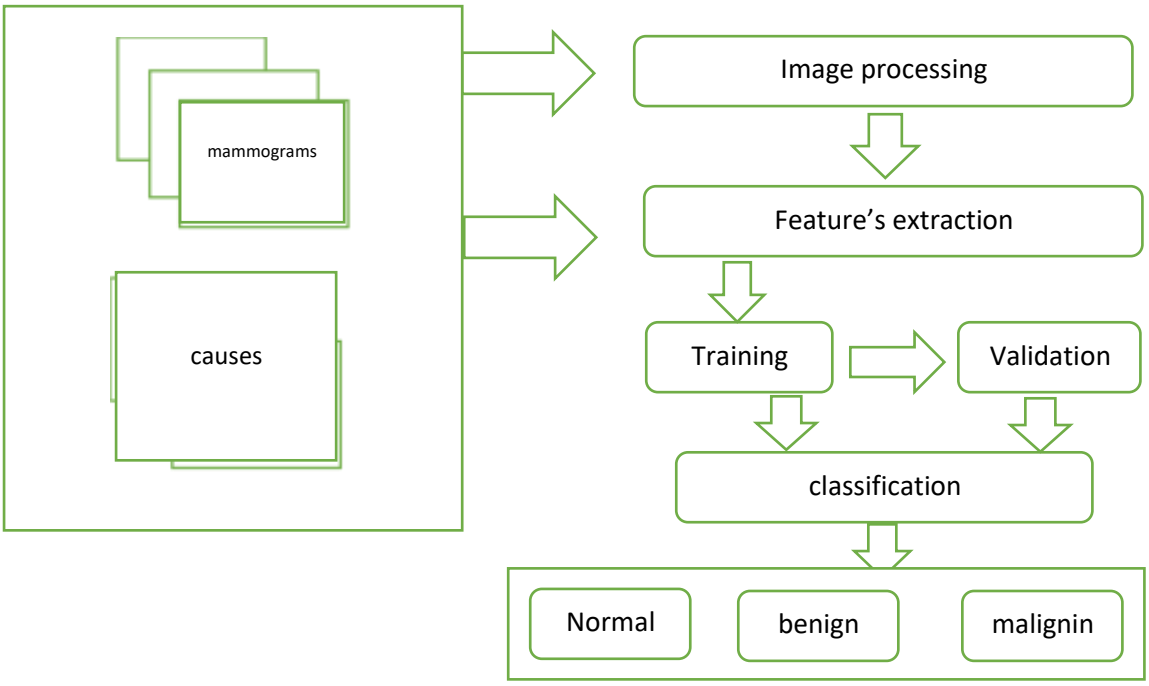
Second year: Implementation of Models of deep learning and Machine learning.

Third year: testing and classification with comparison results obtained.

Fourth year: thesis writing and final submission.

7- Possible outcome

To investigate the efficacy of radiomics in diagnosing patients with breast cancer infected, apply image processing and computer vision processing to improve image quality, preprocessing data of causes and predict using models of deep learning algorithms and help clinicians to get the better results and feedback in diagnosis.



8- Significance of Work / Impact on science/industry/society

Early detection of breast cancer may lead to more precautions to take treatments before the problem exacerbates and expands of tumor, which improves the process of controlling the disease and gave feedback to help the medical specialists.

9- References

[1] M. Hejmadi, Introduction to Cancer Biology, Bookboon,London, UK, 2nd edition, 2010.

[2] American Cancer Society, Breast Cancer Facts and Figures 2017-2018, American Cancer Society, Atlanta, GA, USA, 2017.

[3] American Cancer Society. Cancer Facts & Figures 2021. Atlanta: American Cancer Society; 2021

[4] Kapoor, Pooja Middha, et al. "Assessment of interactions between 205 breast cancer susceptibility loci and 13 established risk factors in relation to breast cancer risk, in the Breast Cancer Association Consortium." *International journal of epidemiology* 49.1 (2020): 216-232.

[5] Hamed, Ghada, et al. "Deep learning in breast cancer detection and classification." *Joint European-US Workshop on Applications of Invariance in Computer Vision*. Springer, Cham, 2020.

[6] Devi, V. Ajantha, and Anand Nayyar. "Fusion of deep learning and image processing techniques for breast cancer diagnosis." *Deep learning for cancer diagnosis*. Springer, Singapore, 2021. 1-25.

[7] Basarslan, M. Sinan, and F. Kayaalp. "Performance evaluation of classification algorithms on diagnosis of breast cancer and skin disease." *Deep Learning for Cancer Diagnosis*. Springer, Singapore, 2021. 27-35.

[8] Patrício, Miguel, et al. "Using Resistin, glucose, age and BMI to predict the presence of breast cancer." *BMC cancer* 18.1 (2018): 1-8.

[9] Zheng, Jing, et al. "Deep learning assisted efficient AdaBoost algorithm for breast cancer detection and early diagnosis." *IEEE Access* 8 (2020): 96946-96954.

[10] Sha, Zijun, Lin Hu, and Babak Daneshvar Rouyendegh. "Deep learning and optimization algorithms for automatic breast cancer detection." *International Journal of Imaging Systems and Technology* 30.2 (2020): 495-506.

[11] Saber, Abeer, et al. "A Novel Deep-Learning Model for Automatic Detection and Classification of Breast Cancer Using the Transfer-Learning Technique." IEEE Access 9 (2021): 71194-71209.

Mr. Aymen Al hejri

Research Scholar

Dr. Archana H. Sable

Research Supervisor